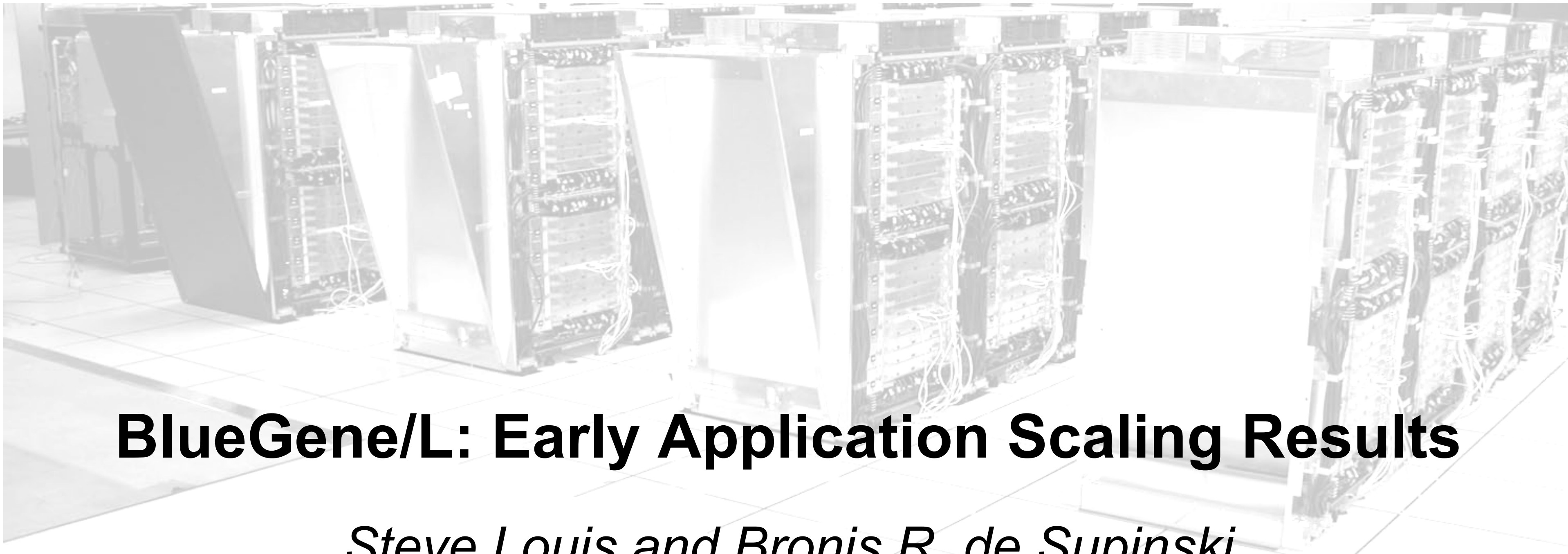


BlueGene/L: Early Application Scaling Results

Steve Louis and Bronis R. de Supinski
Lawrence Livermore National Laboratory

NNSA ASC Principal Investigator Meeting &
BG/L Consortium System Software Workshop
February 24, 2005

UCRL-PRES-209861

This work was performed under the auspices of the U.S. Department of Energy by University of California Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48.

Wide Variety of Early Applications on BlueGene/L

- Blue Matter (IBM) *
- Flash (ANL) *
- Miranda (LLNL) *
- MM5
- Amber7, Amber8
- GAMESS
- QMC (Caltech)
- LJ (Caltech)
- PolyCrystal (Caltech)
- PMEMD (LBL)
- LSMS (ORNL)
- NIWS (NISSEI)
- HOMME (NCAR) *
- Qbox (LLNL)
- ddcMD (LLNL)
- SAGE (LANL)
- SPPM (LLNL)
- UMT2K (LLNL)
- Sweep3d (LANL)
- MDCASK (LLNL)
- GP (LLNL)
- CPMD (IBM/LLNL) *
- TLBE (LBL)
- HPCMW (RIST)
- ParaDiS (LLNL)
- QCD (IBM)*, QCD (BU) *
- NAMD
- PAM-CRASH (ESI)
- Raptor (LLNL) *
- Enzo (SDSC)

Successful scaling tests and science runs completed for key ASC codes

- Using IBM Rochester BG/L hardware through SC2004 (many, many thanks to Jim Sexton of IBM for his help)

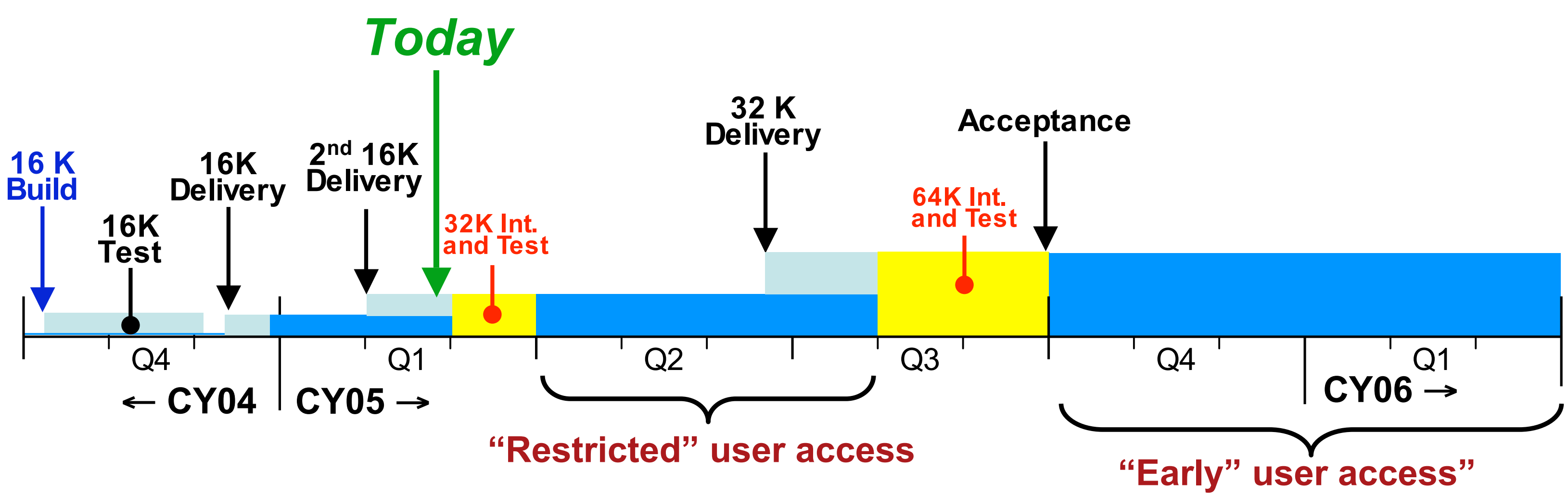
- Several codes running at scale since January 3, 2005 on Livermore's first 16 racks (soon to be 32)

- ASC has concentrated on scaling up the following codes:
  - ddcMD
  - FEQMD
  - GRASP (SNL)
  - hypre/SMG2K
  - LAMMPS (SNL)
  - MDCASK
  - Miranda
  - ParaDiS
  - Qbox
  - Raptor
  - SPaSM (LANL)
  - sPPM (Benchmark)
  - UMT2K (Benchmark)

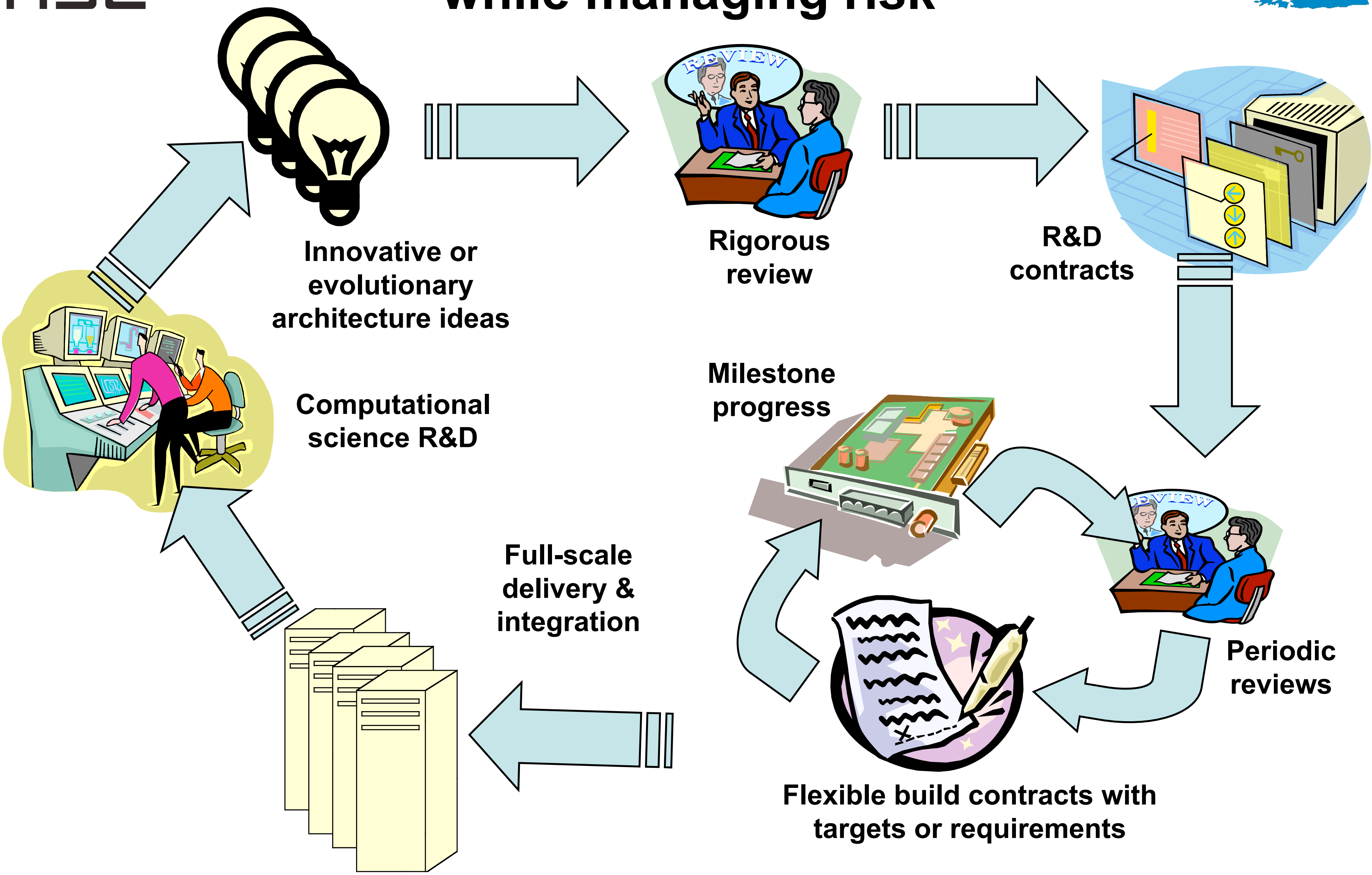
Excellent scaling seen on runs up to 16K nodes and 32K processors

Current Time Line

1. 16 racks delivered in November 2004
   - System tested in December
   - Runs started in January
2. 16 additional rack delivery early February
   - 32-rack runs start in April
3. 32 additional rack delivery in mid 2005
   - integration to 64 rack system
   - Machine shake out and test
   - 64-rack scaling/science follows
4. Late 2005 expanded early user access and more science runs

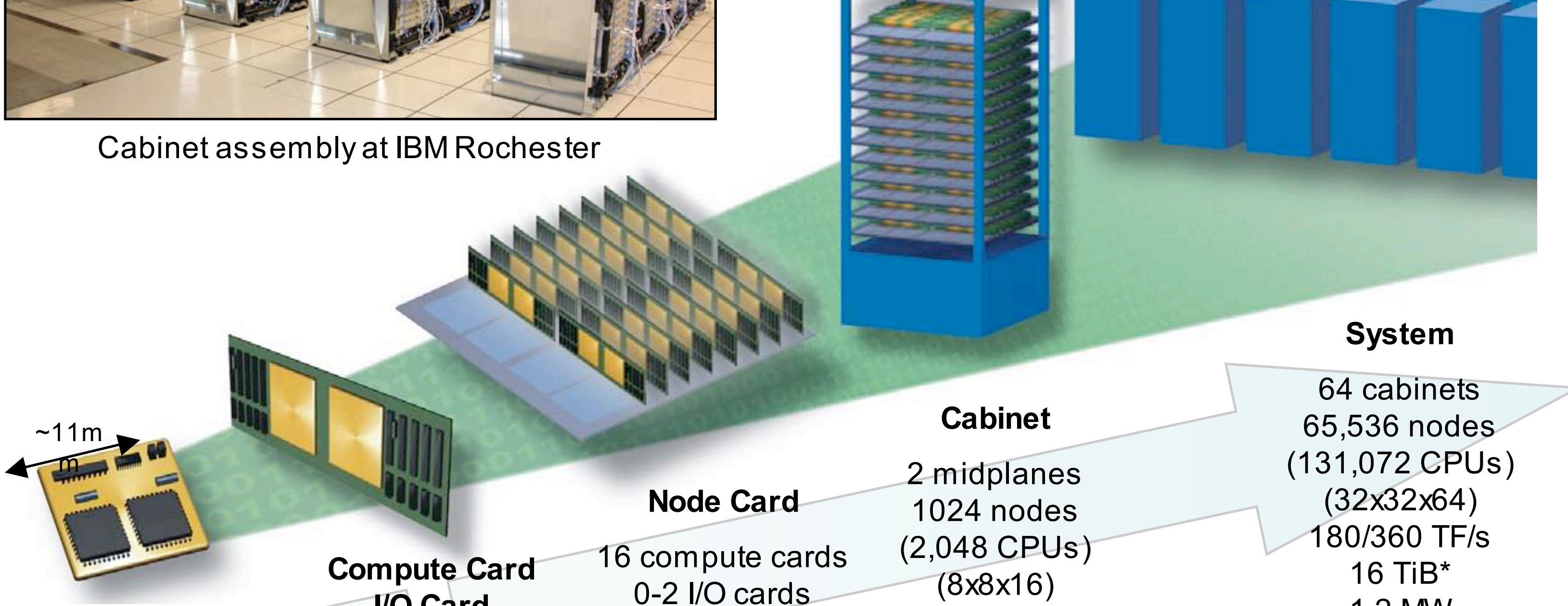


DOE strategic investment in ASC advanced architecture encourages innovative ideas while managing risk



BlueGene/L scales to 360 TF with modified COTS and custom parts

Cabinet assembly at IBM Rochester



Compute Chip

2 processors
2.8/5.6 GF/s
4 MiB* eDRAM

(compare this with a 1988 Cray YMP/8 at 2.7 GF/s)

Compute Card
I/O Card

FRU (field replaceable unit)
25mmx32mm
2 nodes (4 CPUs)
(2x1x1)
2.8/5.6 GF/s
256/512 MiB* DDR

Node Card

16 compute cards
0-2 I/O cards
32 nodes
(64 CPUs)
(4x4x2)
90/180 GF/s
8 GiB* DDR

Cabinet

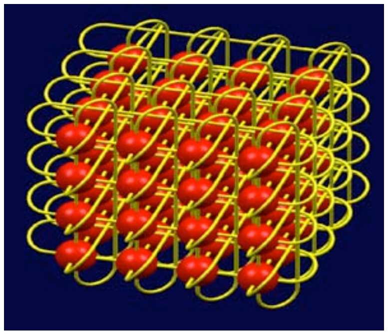
2 midplanes
1024 nodes
(2,048 CPUs)
(8x8x16)
2.9/5.7 TF/s
256 GiB* DDR
15-20 kW

System

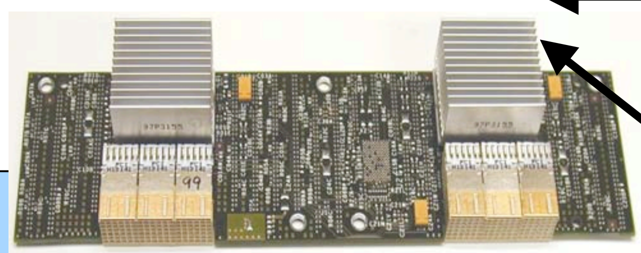
64 cabinets
65,536 nodes
(131,072 CPUs)
(32x32x64)
180/360 TF/s
16 TiB*
1.2 MW
2,500 sq.ft.
MTBF 6.16 Days

* http://physics.nist.gov/cuu/Units/binary.html

Architectural features promote efficiency and scaling for important applications

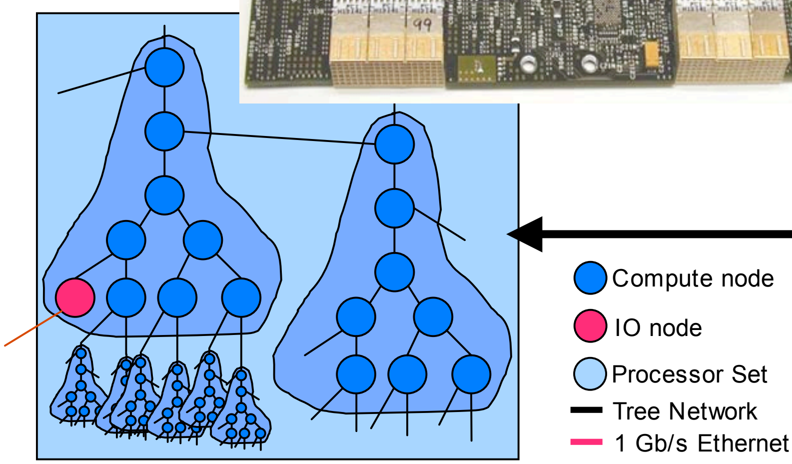


- **Multiple complementary interconnects support diverse application scaling requirements**
  - 3D torus with bi-directional nearest-neighbor links
  - 2.1 GB/s combining tree for fast global reductions
  - Low-latency global barrier network



- **High reliability expected from high level of integration using system-on-a-chip technology**

- **Architectural enhancements improve single node performance**



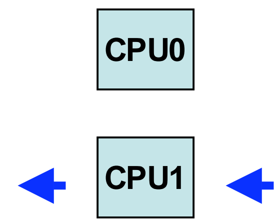
- **Software architected with very powerful "divide and conquer" technique for software scale-up**

The BG/L project is a focused effort to enable important science and to lead the way to cost-effective petaFLOP/s computing

Two ways for apps to use hardware

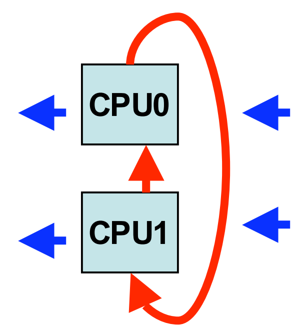
Mode 1 (Co-processor mode - CPM):

- CPU0 does all the computations
- CPU1 does the communications
- Communication overlap with computation
- Peak comp perf is 5.6/2 = 2.8 GFlops

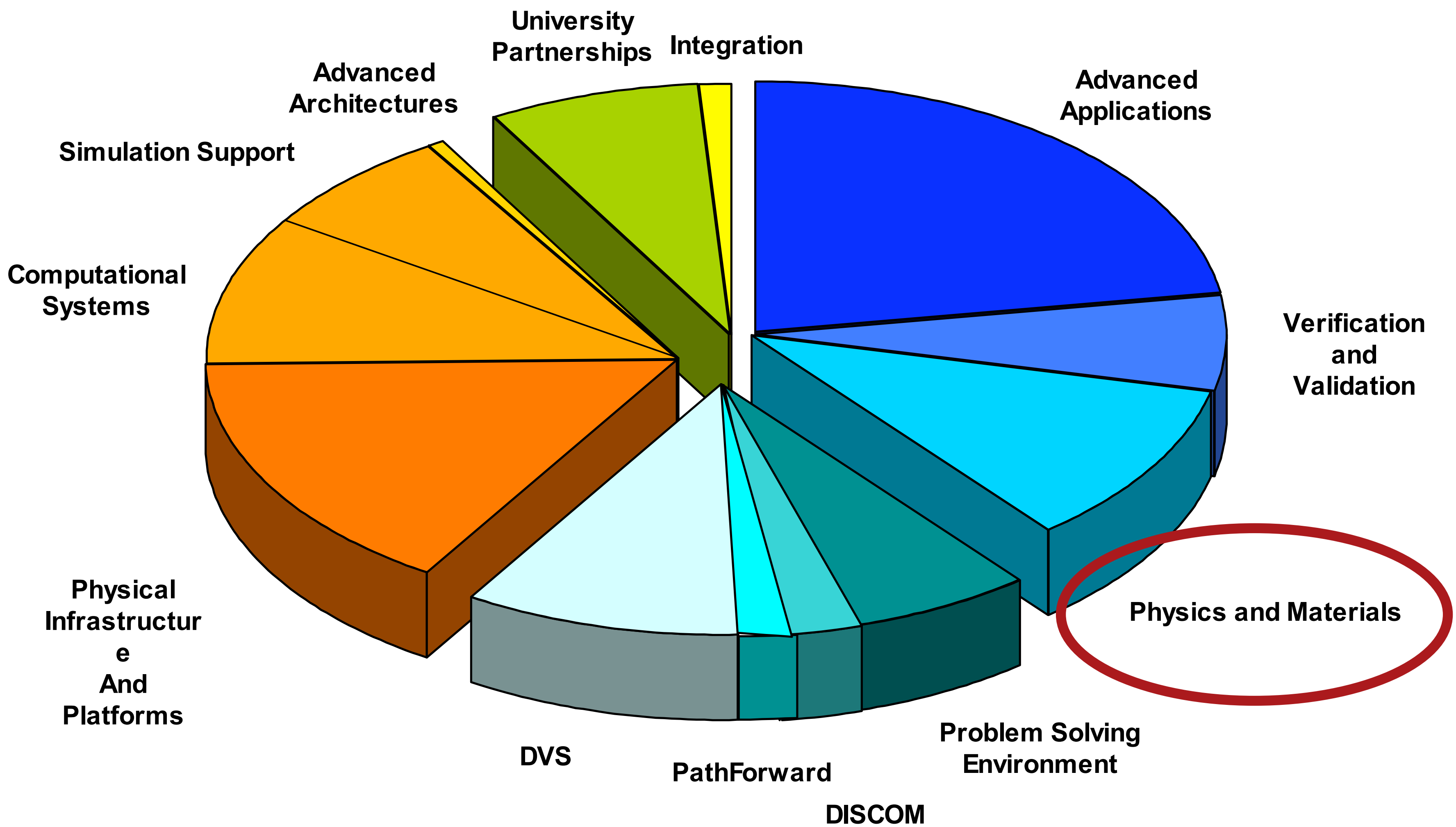


Mode 2 (Virtual node mode - VNM):

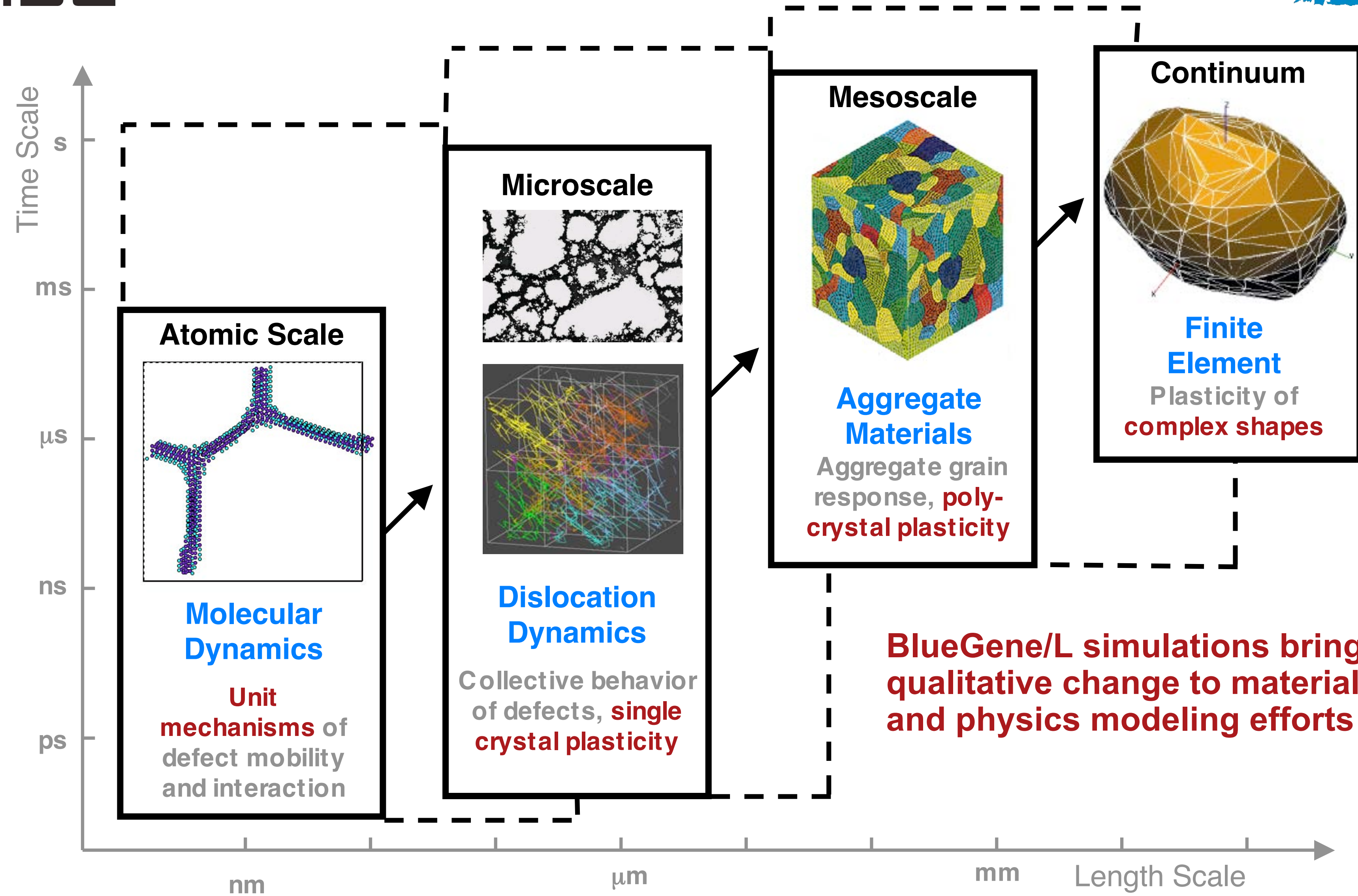
- CPU0, CPU1 independent “virtual tasks”
- Each does own computation and communication
- The two CPU’s talk via memory buffers
- Computation and communication cannot overlap
- Peak compute performance is 5.6 GFlops



Physics and Materials is initial ASC focus for the early BlueGene/L apps



BlueGene/L will allow overlapping evaluation of models for first time



Predicting the mechanical strength of material from first principles is difficult

- *In fact, it has remained a grand challenge for computational materials science for several decades…*

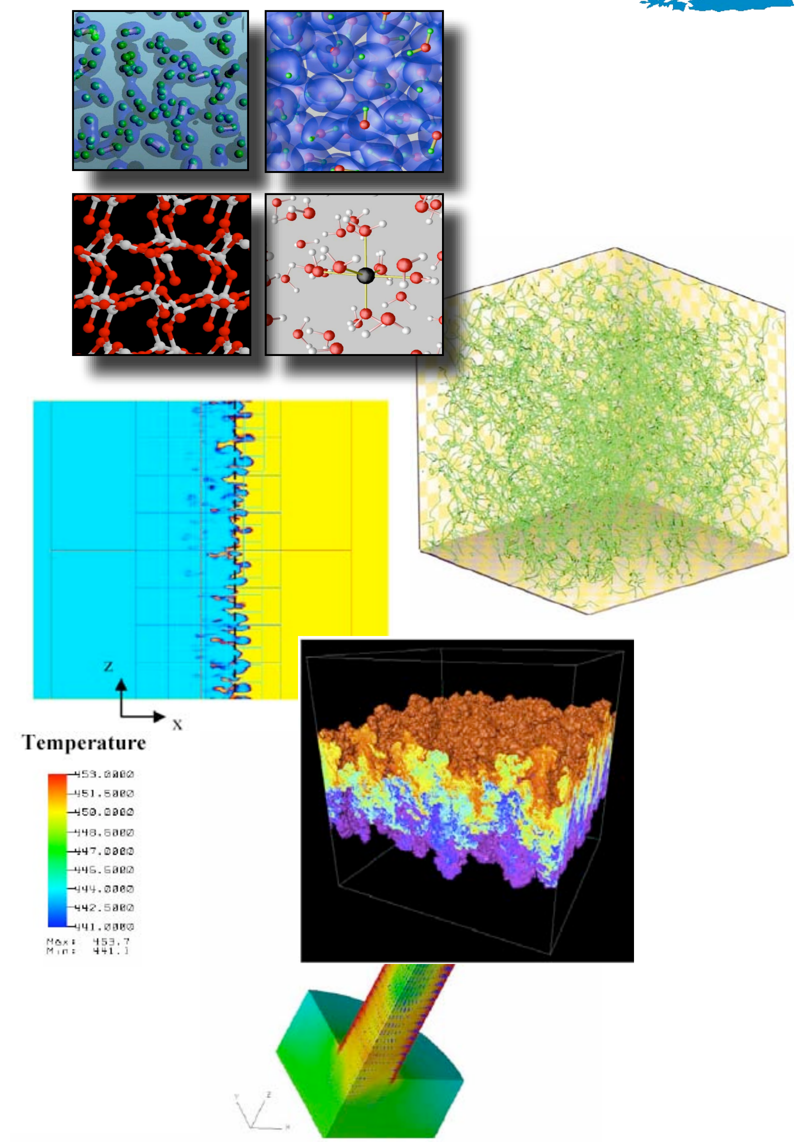
- **At a fundamental scale, we need to consider the atomistic structure of a "dislocation"** and the atomistic mechanisms of dislocation motion.

- **At higher scale, we need to consider interaction of many dislocations that form complex patterns on a micron scale** to understand the plastic strength of a single crystal.

- **Single-crystal strength is used to model a poly-crystal of many crystal grains,** which in turn supports yet higher-scale finite-element models of complex-shape object deformation.

Criteria for “First-Wave” Applications

- **First-wave target ASC applications**
  - **Efforts identified to be ready for programmatic science runs with early machine availability, and ongoing assessment of code suitability**
- **Assessment criteria for early apps**
  - **Importance to the ASC program**
  - **Enthusiasm within the code group**
  - **Potential for good code scaling *(i.e., simpler architectural needs)***
- **Second-wave target ASC apps are also being identified using similar criteria to those above**



Also examining many other science applications to run on BlueGene/L

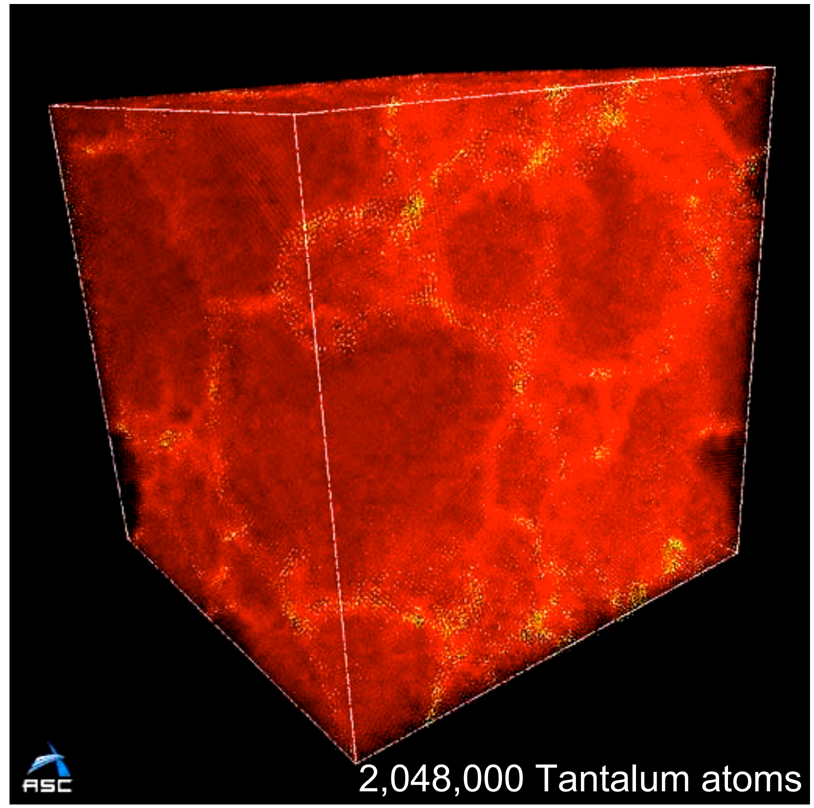
- **ALPS** – Predictive modeling of laser plasma interaction
- **AMRh** – High Reynolds-number fluid flows
- **BOUT** – MFE boundary-layer turbulence
- **DJEHUTY** – 3D modeling of stars
- **DYNA3D** – Structural mechanics
- **EMSOLVE** – Electromagnetic coupling with structures
- **FMC** – Solves Schroedinger Equation for a many-fermion system
- **GFMD** – Greens-function molecular dynamics
- **HYDRA** – Implicit radiation diffusion solver, multi-mode instabilities
- **IRS** – Implicit radiation solver
- **ParaDyn** - LLNL parallel engineering code based on DYNA3D
- **PF3D/Z3** – Predictive modeling of laser plasma interactions
- **ROLEX** – Detailed-accounting opacity
- **SAGE** – LANL widely used adaptive-grid Eulerian hydrodynamics

Classical MD – ddcMD: Rapid resolidification in tantalum

• Scalable, general purpose code for performing classical molecular dynamics (MD) simulations using highly accurate MGPT potentials

• MGPT semi-empirical potentials, based on a rigorous expansion of many body terms in the total energy, are needed in to quantitatively investigate dynamic behavior of transitions metals and actinides under extreme conditions

64K and 256K atom simulations on 2K nodes are order of magnitude larger than previously attempted; based on 2M atom simulation on 16K nodes, *close to perfect* scaling expected for full machine ("very impressive machine" says PI…)



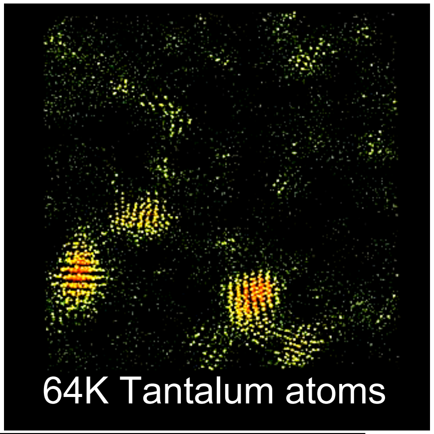
Visualization of important new scientific findings already achieved on BG/L: Molten Ta at 5000K demonstrates solidification during isothermal compression to 250 GPa

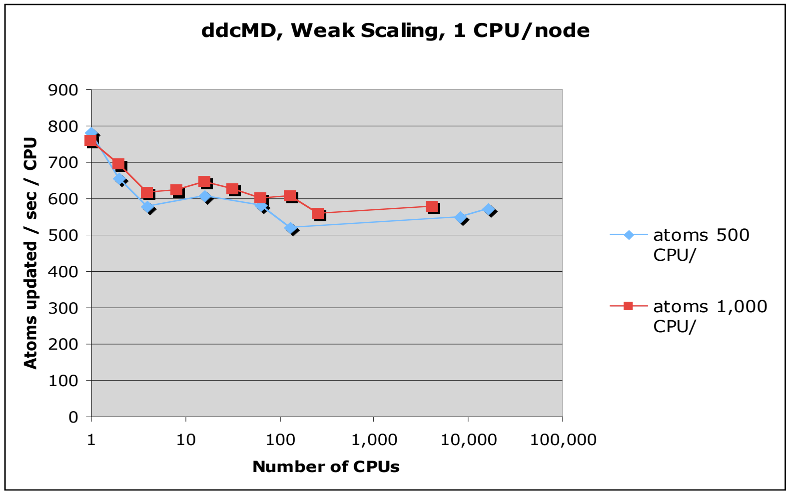
Excellent scaling of ddcMD on BG/L supports greater understanding of resolidification process

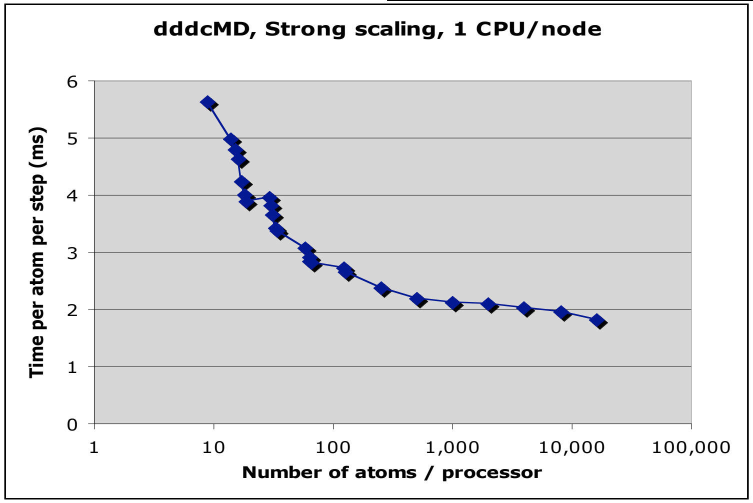
• Nucleation of solid is initiated at multiple independent sites throughout each sample cell

• Growth of solid grains initiates independently, but soon leads to grain boundaries which span the simulation cell: size of cell is now influencing continued growth

•2,048,000 simulation recently performed indicates formation of many more grains







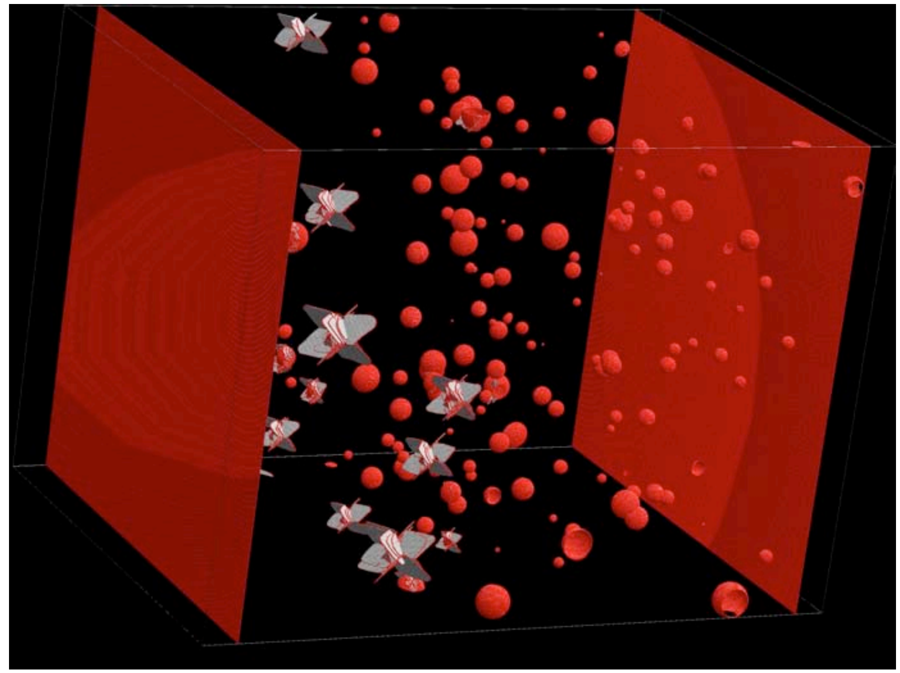
ddcMD is already using 32K* CPUs of BG/L for unprecedented multi-million atom MGPT simulations

*** Virtual node mode gives 1.9x performance out to 1K processors**

Classical MD - SPaSM

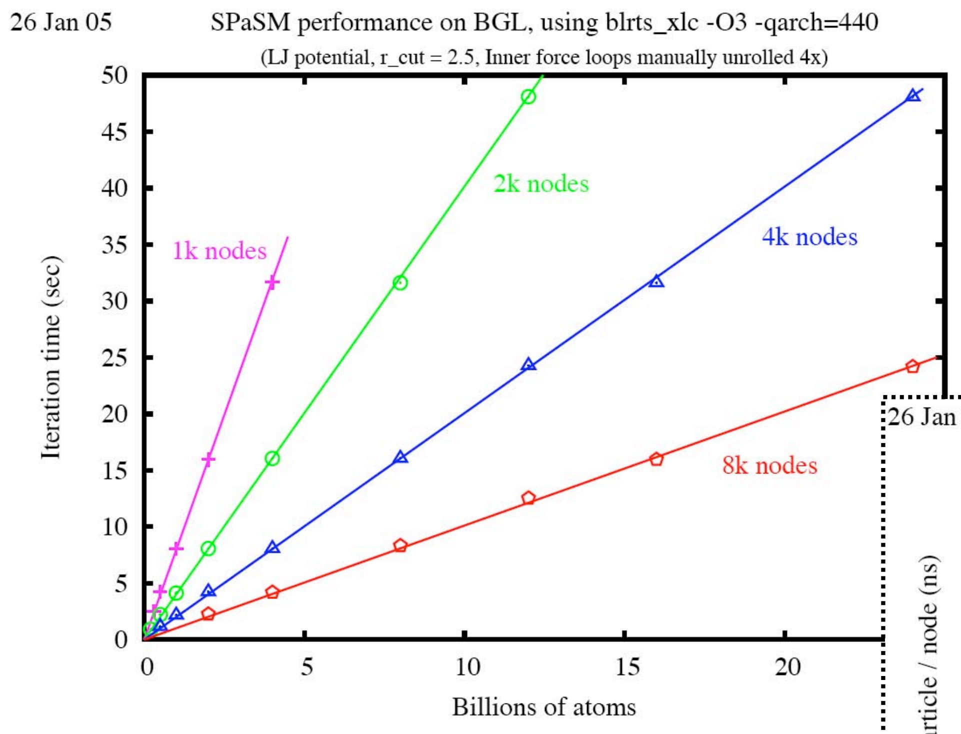
- A high performance (1993 and 1998 Gordon Bell prizes) code for **S**calable **Pa**rallel **S**hort-range **M**olecular dynamics simulations

- A variety of finite-range empirical potentials are implemented, including EAM and MEAM for metals, Stillinger-Weber Si/Ge, and a reactive empirical bond-order (REBO) potential for detonation studies.



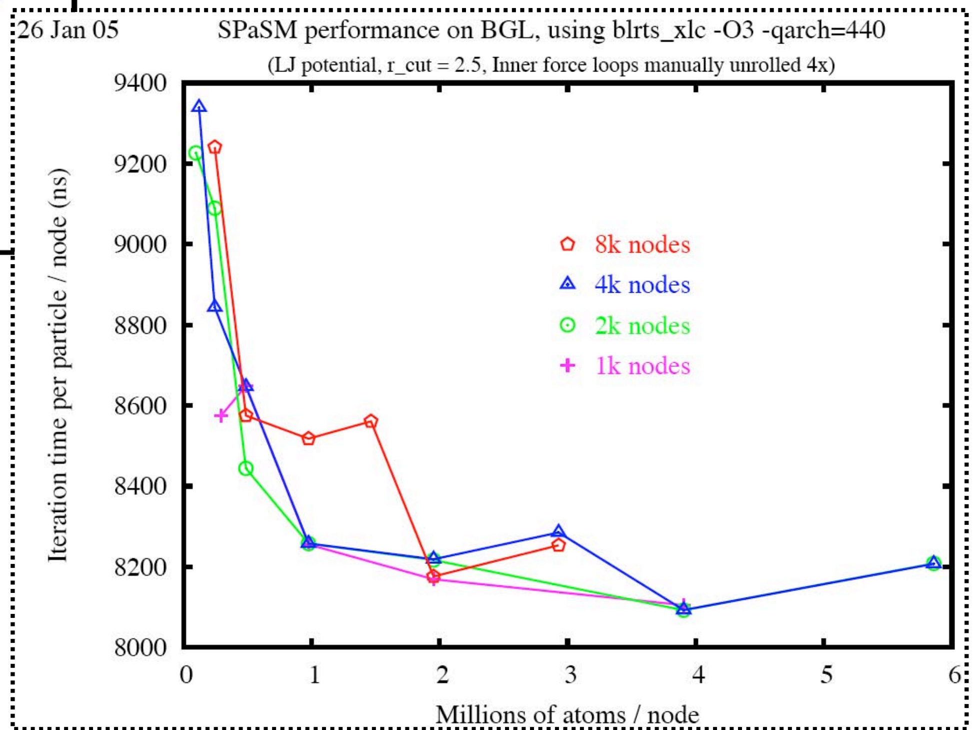
SPaSM has exhibited excellent scaling for up to 100 billion atoms on 16,384 nodes, and an initial production run on 8k nodes simulated the shock loading of a 2.1 billion atom EAM copper crystal with 0.41% (by volume) voids. BG/L will enable the exploration of an entirely new class of (previously intractable) problems such as this.

SPaSM performance on BG/L with 100,000 — 6,000,000 atoms per node



- Excellent scaling is seen for >1 million atoms per node

- This will extend to smaller sizes for more complex potentials, which typically have much higher CPU/communication ratios



GRASP: Scalable Molecular Dynamics Code for Reactive Force Fields (SNL)

• A scalable, general purpose code for performing classical molecular dynamics (MD) simulations

• Supports a wide range of different force fields: twobody, threebody, Tersoff, EAM, ReaxFF, electrostatics, charge equilibration

•Applications include: Radiation Damage, Materials Interfaces, Explosives, MEMS

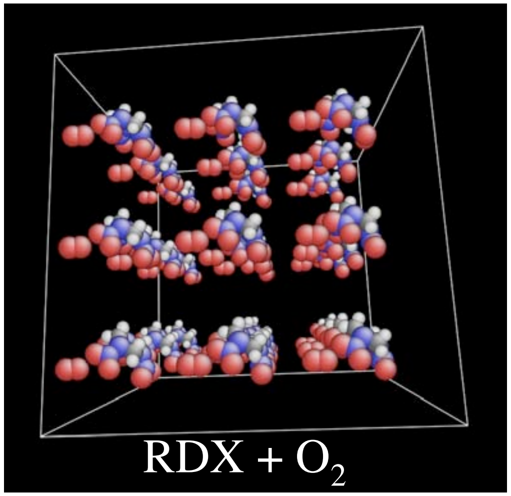
•Standard version ported to BGL without difficulties

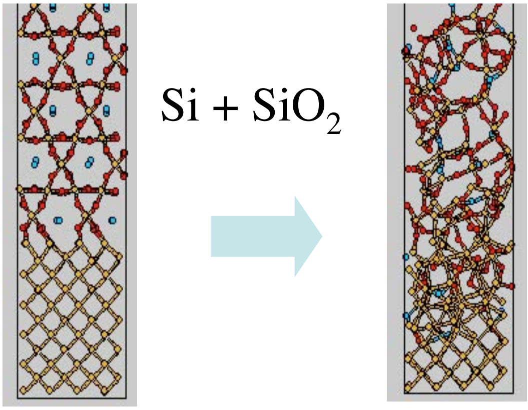
•Development version combining C++ and Fortran implements ReaxFF force field. Used BGL to test the code. Several software bugs detected and fixed.

•Absolute speed for ReaxFF close to 3 GHz Pentium cluster.

•Stillinger-Weber silicon benchmark scales well

- 70% efficiency at 122 atoms/CPU on 16k processors.
- 35% efficiency at only 4 atoms/CPU on 4k processors.





GRASP performance on BG/L

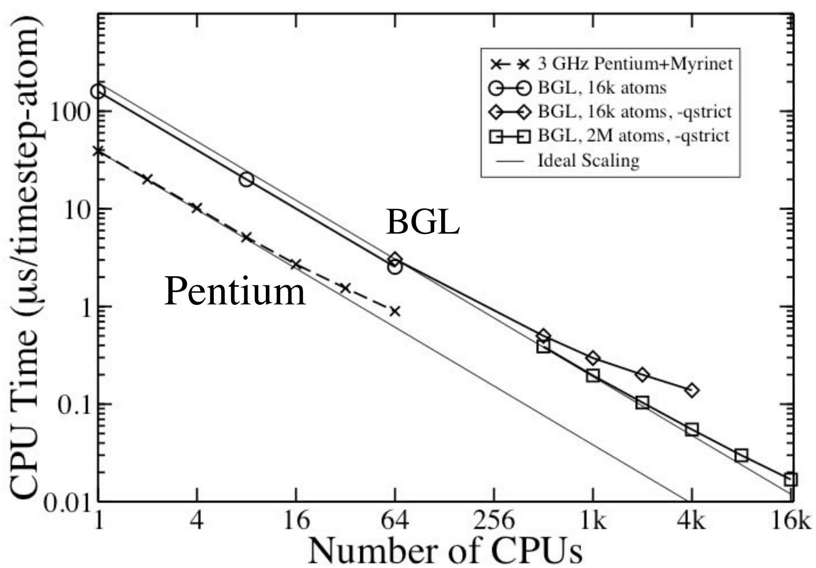
and comparison to HP cluster (3 GHz Pentium+Myrinet

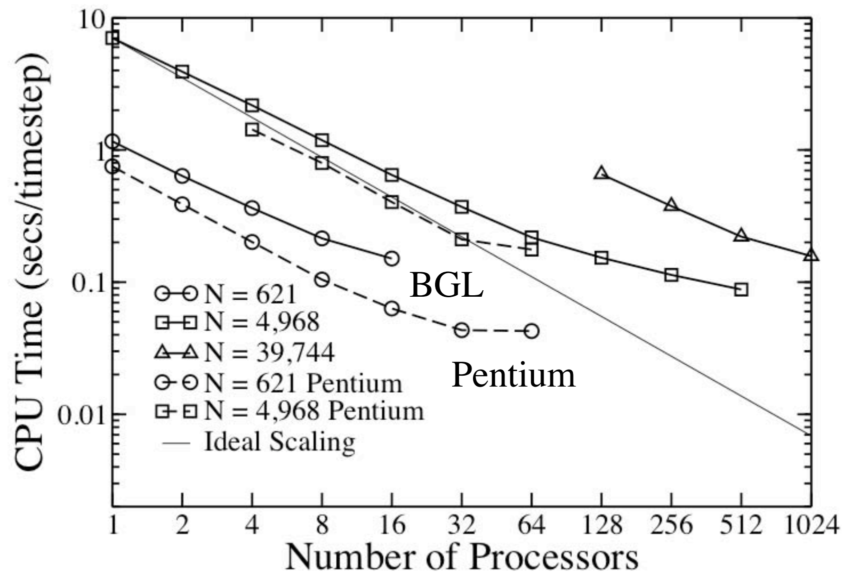
α-Silicon crystal

Rhombohedral periodic cell
Stillinger-Weber force field

RDX Explosive with Oxygen

ReaxFF force field with charge equilibration



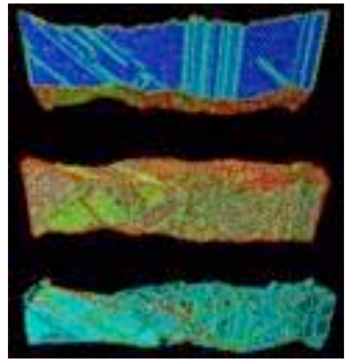


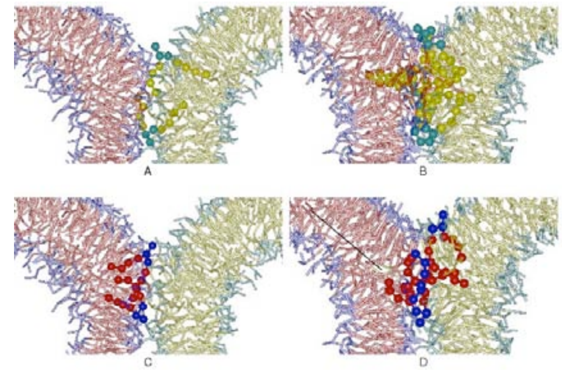
LAMMPS Classical MD (SNL)

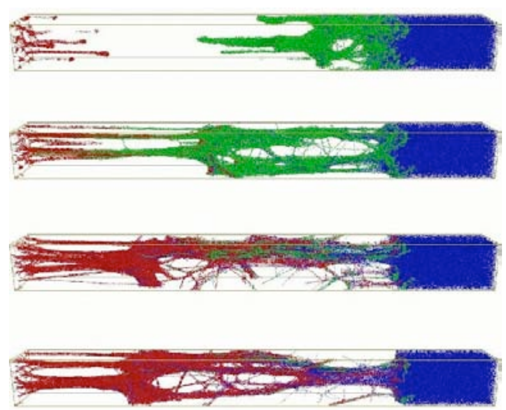
- LAMMPS = Large-scale Atomic/ Molecular Massively Parallel Simulator

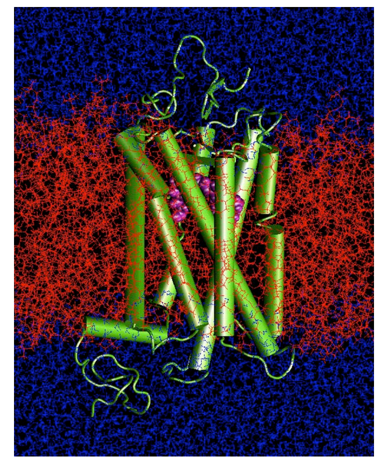
- LAMMPS is a classical molecular dynamics code that models an ensemble of particles in a liquid, solid, or gaseous state. It can model atomic, polymeric, biological, metallic, or granular systems using a variety of force fields and boundary conditions.

- On parallel machines, LAMMPS uses spatial-decomposition techniques to partition the simulation domain into small 3d sub-domains, one of which is assigned to each processor.



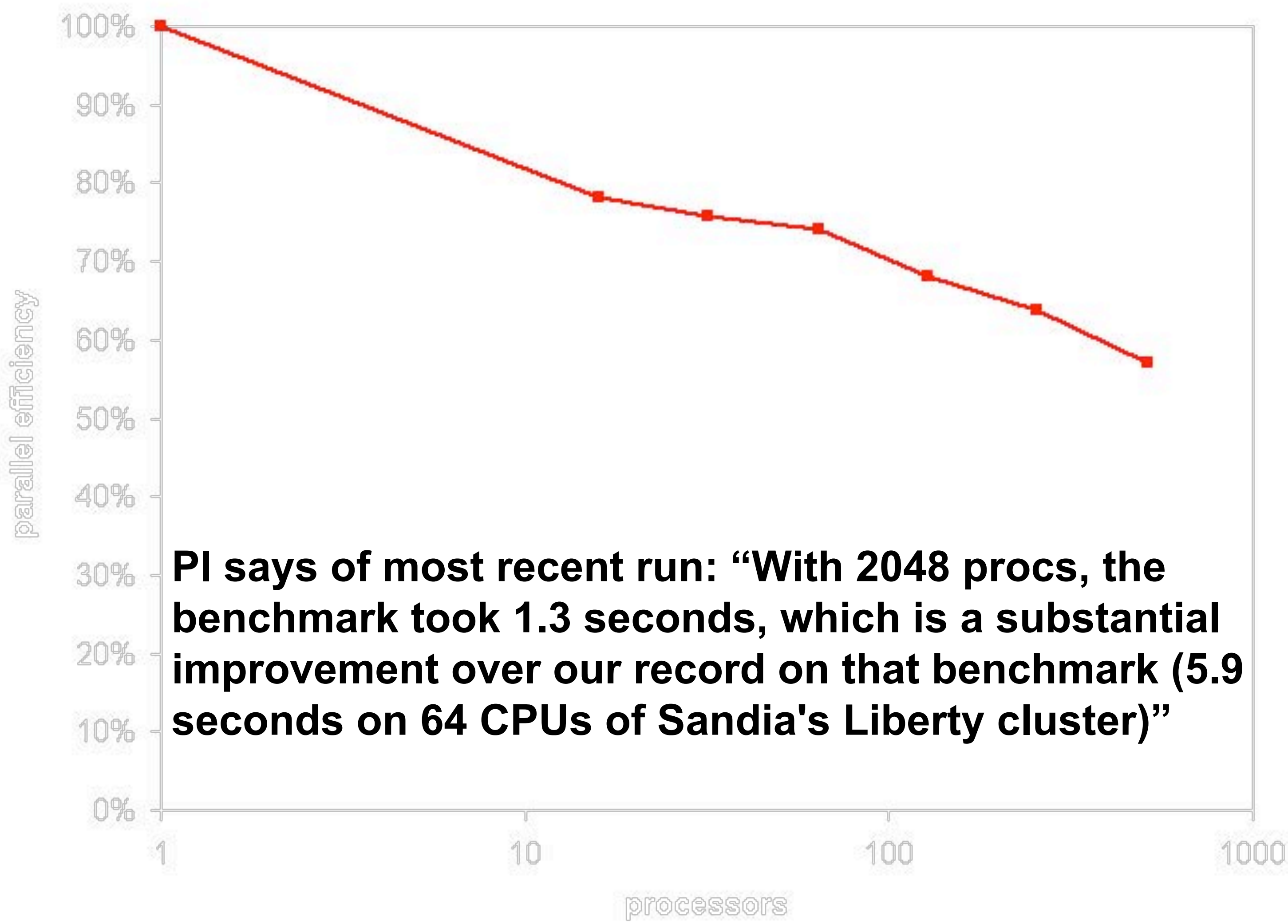






LAMMPS has been tested on up to 512 BG/L processors so far, and shown good scaling on a fixed-size (32,000 atoms) problem (strong scaling).

LAMMPS strong scaling on BG/L with 32,000 atom fixed size problem



Classical MD - MDCASK

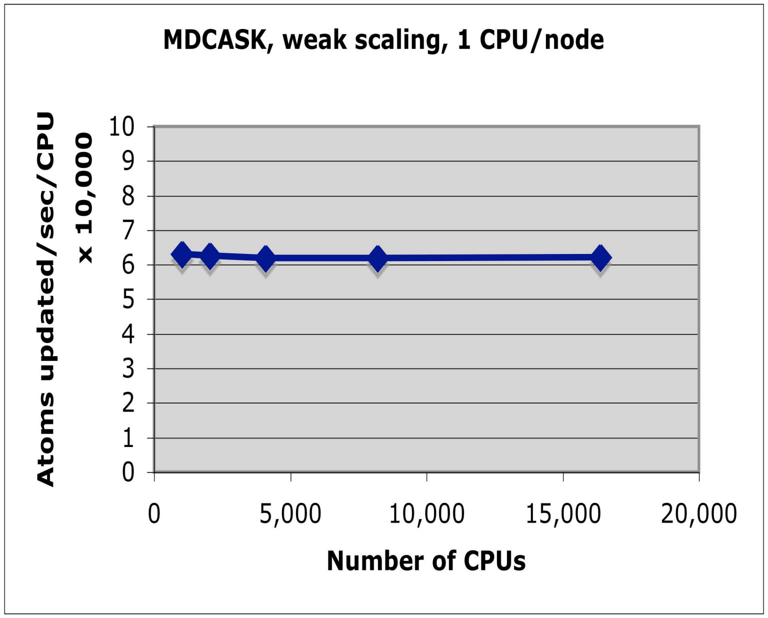
- MDCASK simulates the motion of large collections of individual atoms using the classical laws of Newtonian mechanics and electrostatics.
- Capable of using a wide variety of inter-atomic potentials allowing simulation of metals, semiconductors, insulators, and glasses

• Weak scaling (with a constant 250,000 atoms per processor) was tested up to 16,384 processors with excellent results.

• Virtual node mode yields a factor of 1.78 speedup.

• To simulate 1 ns with 10^{10} atoms requires ~ 8 days on the full-sized BG/L.

• Strong scaling tests perform well down to ~2,000 atoms / node.



MDCASK is ready to apply the full power of BGL to multi-billion atom simulations.

First-Principles MD - Qbox

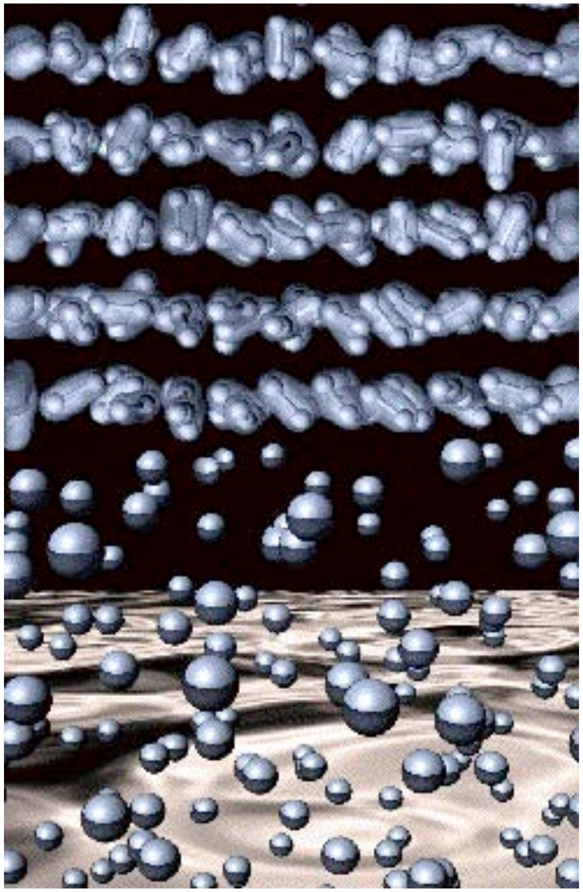
•Qbox is a C++/MPI implementation of the plane-wave, pseudopotential, ab initio molecular dynamics method within Density Functional Theory (DFT). It is developed at LLNL.

•Massively parallel C++ / MPI implementation with specialized 3D FFTs

•Routinely used at LLNL for simulations of condensed matter subjected to extreme such as high pressure and high temperature, as well as in nanotechnology and biochemistry applications.

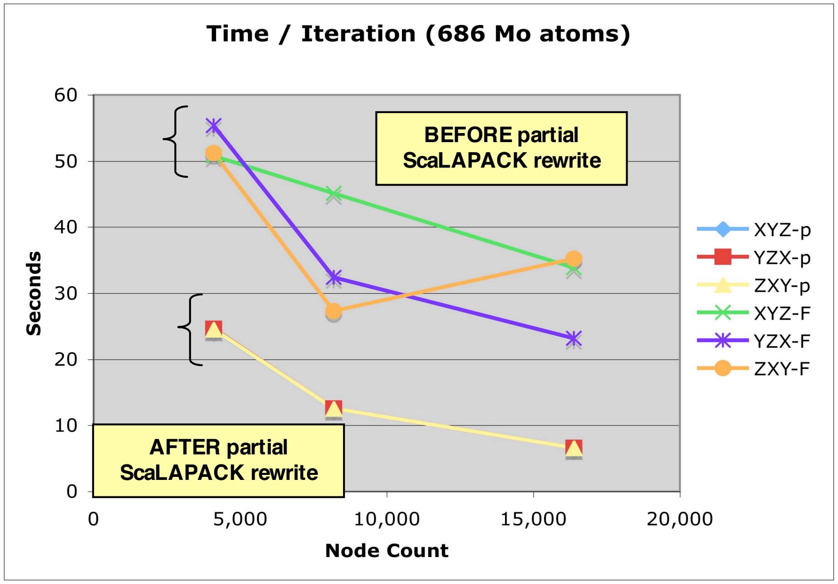
• 686-atom Mo solid and other heavy metal simulations are under way

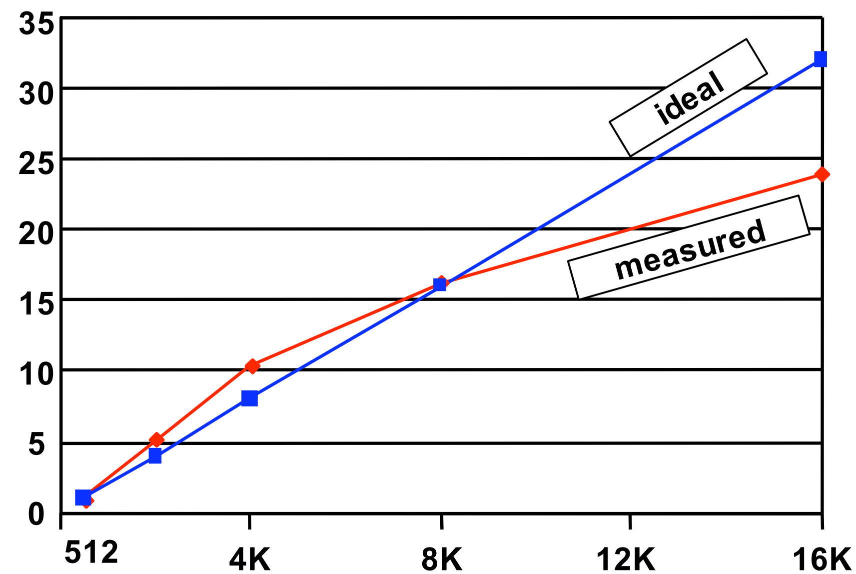
•Scalability tests on BG/L show that Qbox can achieve a 3x speedup when solving a given problem on 16384 nodes instead of 4096 nodes. This represents a 75% parallel efficiency. Further optimizations will provide even greater efficiency.



This figure (generated with GP, pre-cursor of Qbox) was recently used as the cover of the October 7, 2004 issue of the journal Nature.

Qbox: (strong) scaling on BG/L
Solid Molybdenum simulation

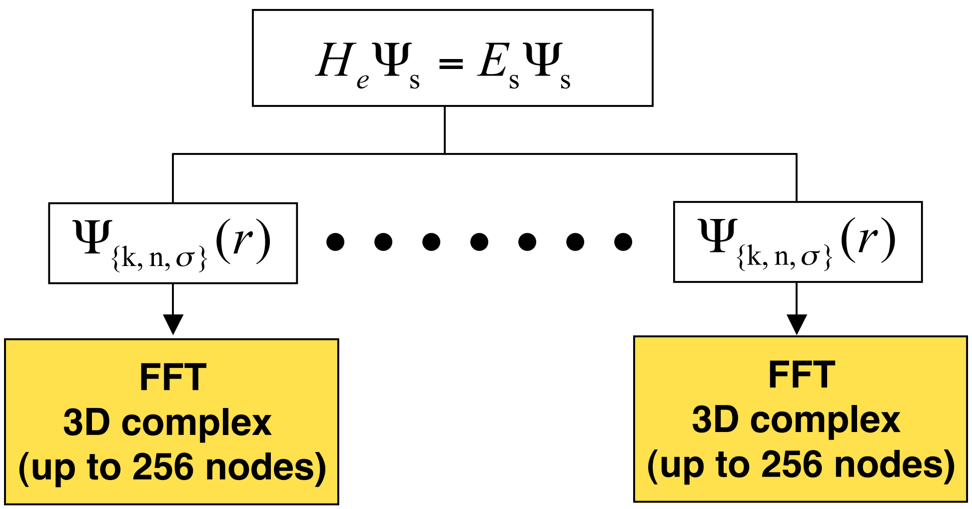




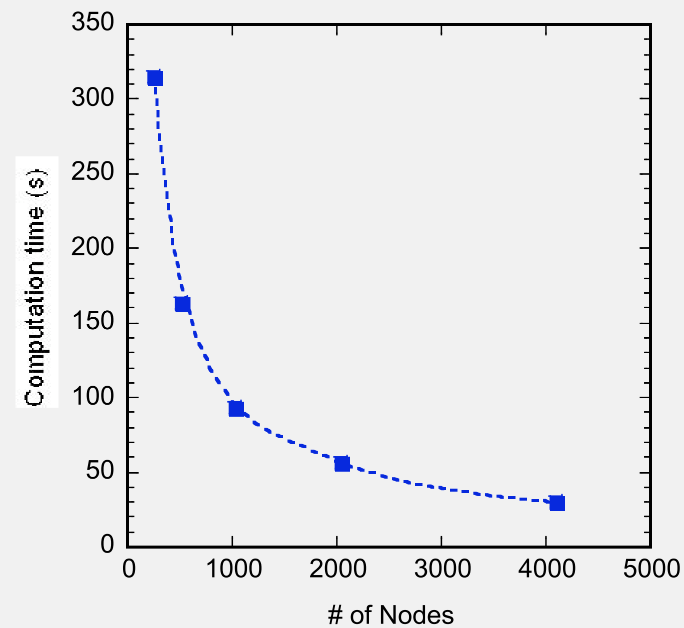
- Some lessons learned:
  - Node mapping is critical, can result in a 2x speedup
  - Using two CPUs / node yields ~1.5x speedup
  - Mixed “AIX/Linux” development environment, w/evolving compilers and nascent MPICH-2 BG/L device, has proved challenging
  - 16k task algorithm scaling frequently requires modifications: Rewrote some ScaLAPACK functions to improve scaling above 4k nodes
- Current efforts target generating efficient node mappings, optimization of linear algebra operations and parallel I/O

First-Principles MD - FEQMD

- **FEQMD has required a complete "inversion" of its parallelization strategy.**
- **Data arrangement: {X, Y, Z} in real space, {Z, Y, X} in Fourier space reduces the number of transpose operations.**
- **Further scaling and optimization work is underway.**



Runs on BG/L for 64 atoms, one self-consistent calculation without I/O



BGL is ~25x the power of ASCI Q, where we currently simulate 128 atoms

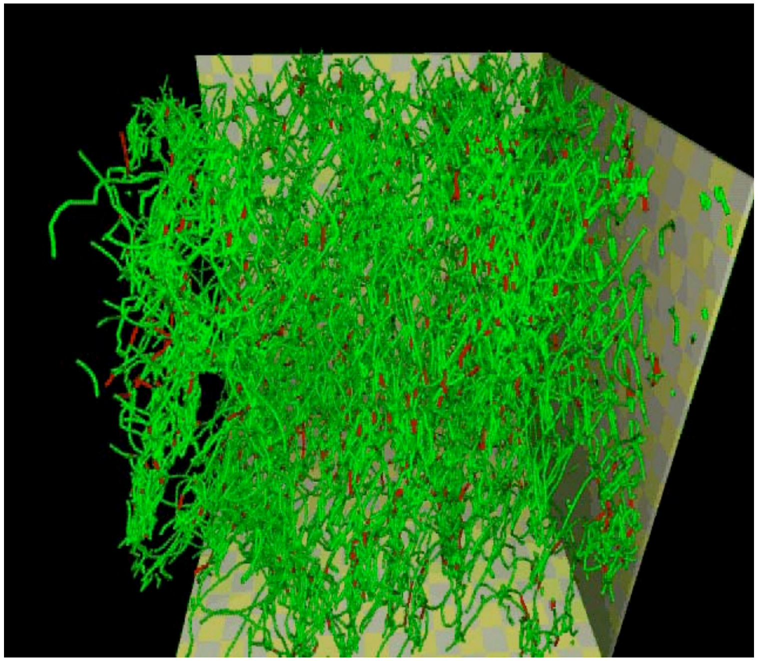
Dislocation Dynamics - ParaDiS (**Para**llel **Di**slocation **S**imulator)

• New LLNL code for direct computation of plastic strength of materials

• Tracks simultaneous motion of millions of dislocation lines

• Promises to close the computational performance gap that prevents scientists from understanding the fundamental nature of material strengthening (or hardening)

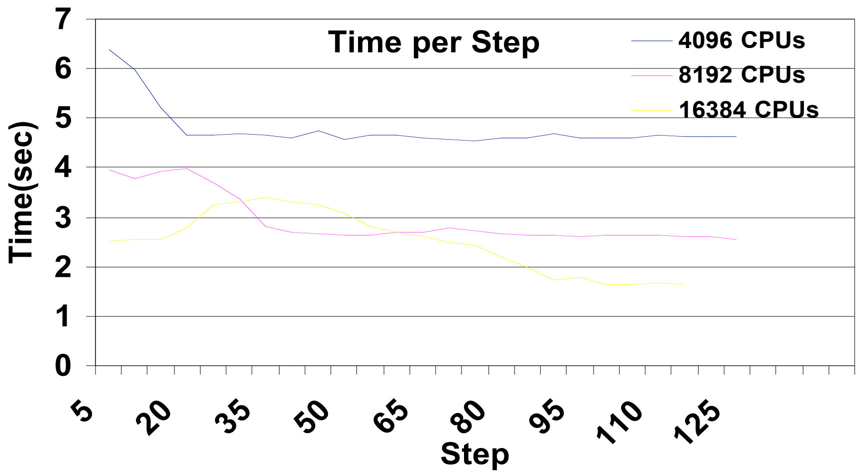
ParaDiS has run on 16,384 nodes of BG/L, and is currently investigating scaling and dynamic load balancing issues to achieve higher efficiencies.

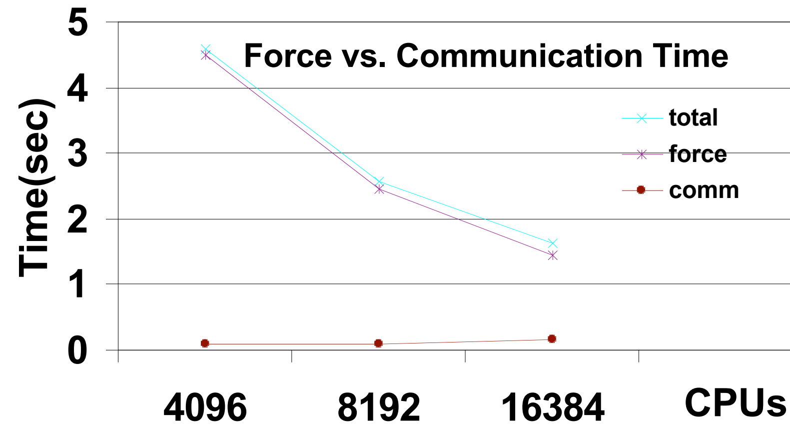


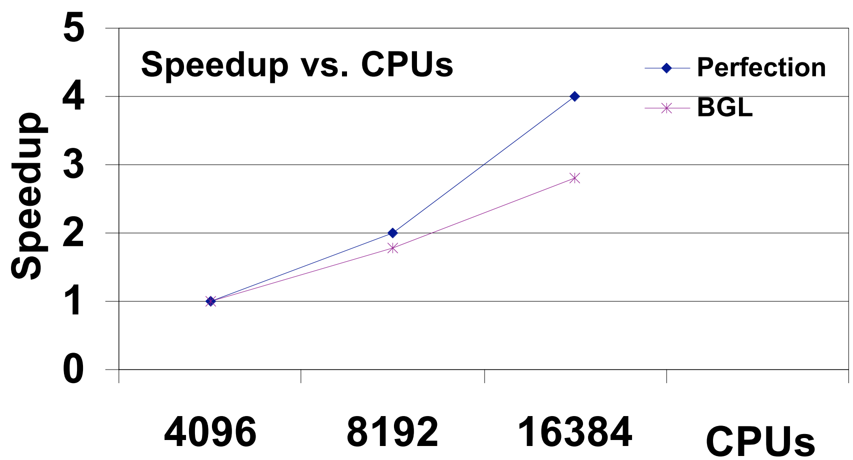
Killer applications:

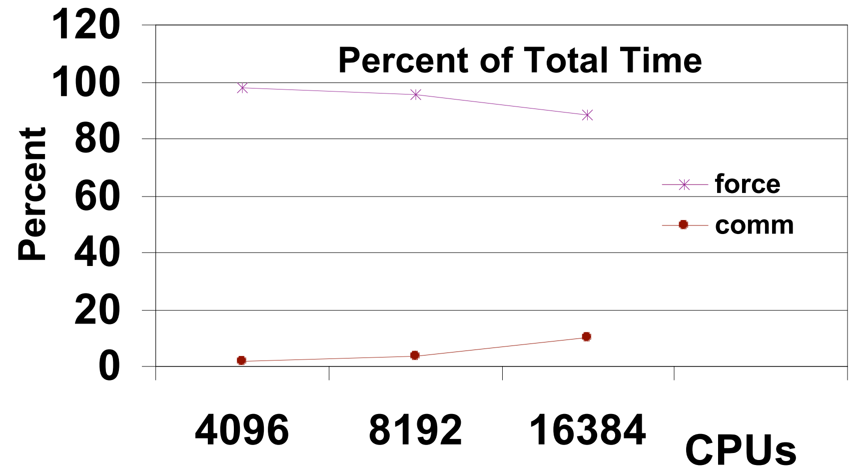
- full simulation of poly-crystal solidification from melt
- alloy microstructure evolution during plastic deformation
- science of ultra-fast polymer crystallization

Dynamic load balancing key to scaling ParaDis to large node counts



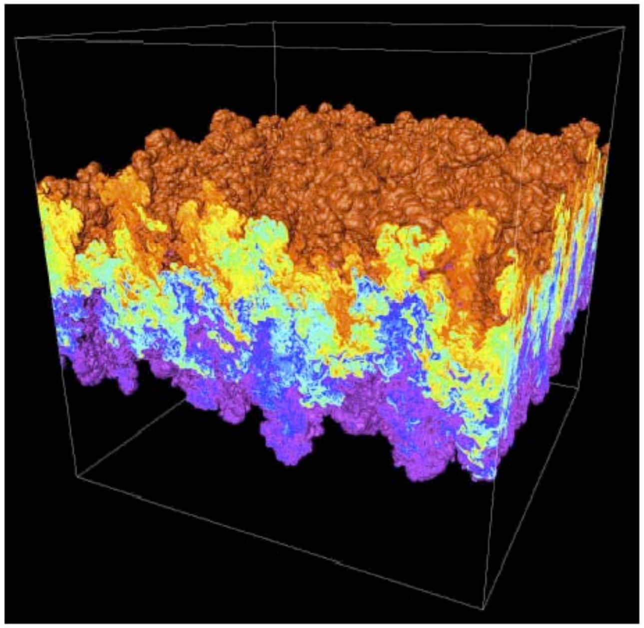






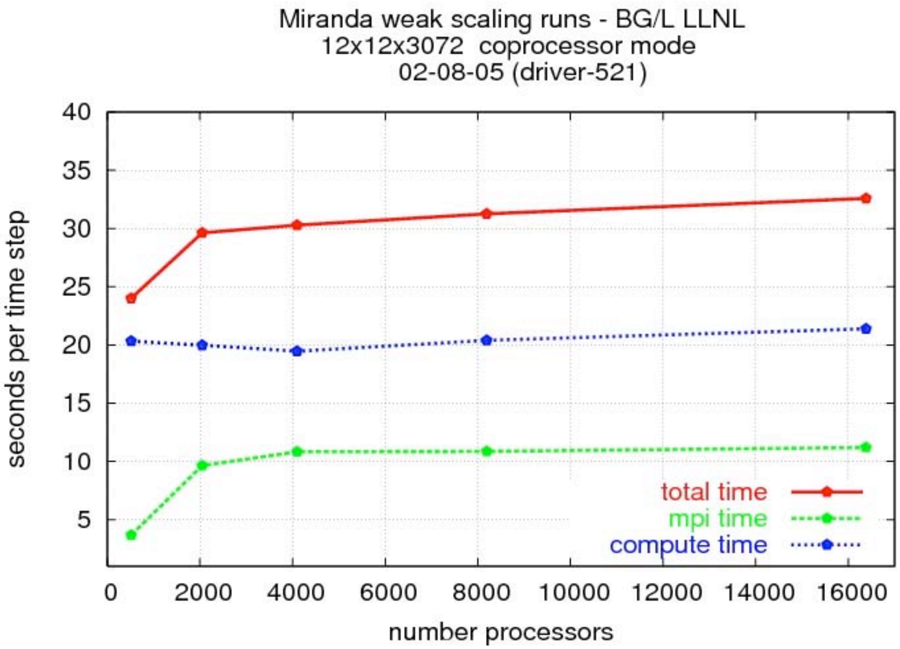
Instability and Turbulence - Miranda

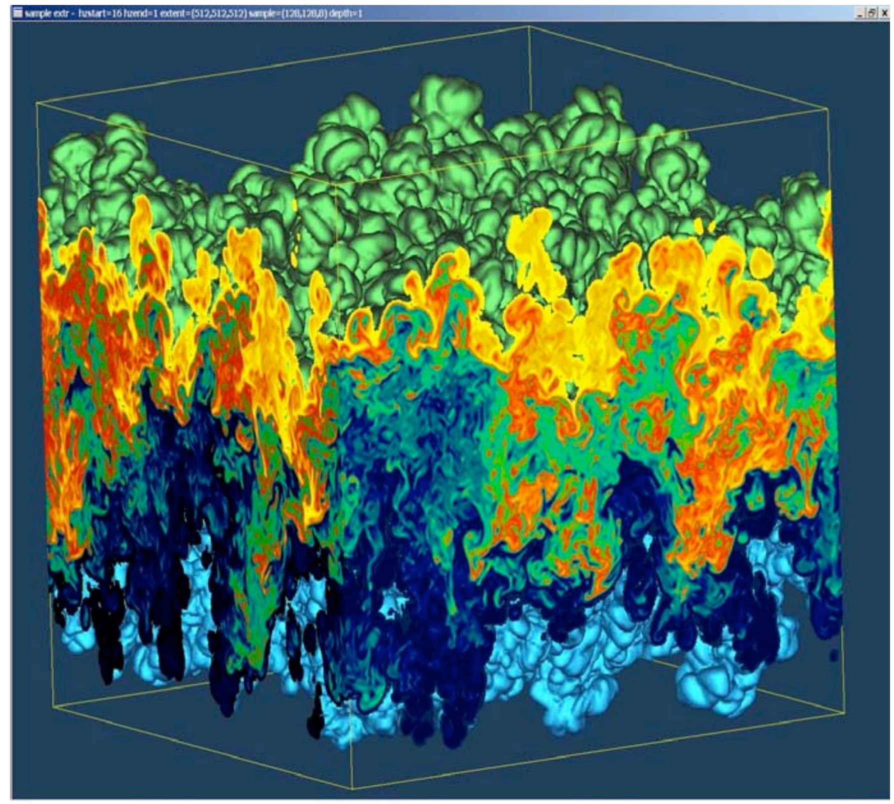
- High order hydrodynamics code for computing fluid instabilities and turbulent mix
- Employs FFTs and band-diagonal matrix solvers to compute spectrally-accurate derivatives, combined with high-order integration methods for time advancement
- Contains solvers for both compressible and incompressible flows
- Has been used primarily for studying Rayleigh-Taylor (R-T) and Richtmyer-Meshkov (R-M) instabilities, which occur in supernovae and Inertial Confinement Fusion (ICF)



Miranda has successfully run on 16,384 nodes on BG/L and also on 32,768 processors in "virtual node" mode. BG/L enables wide range of scales in space and time necessary to represent turbulent flows of interest. Good time-to-solution improvement from MCR to BG/L .

Miranda Weak Scaling on BG/L



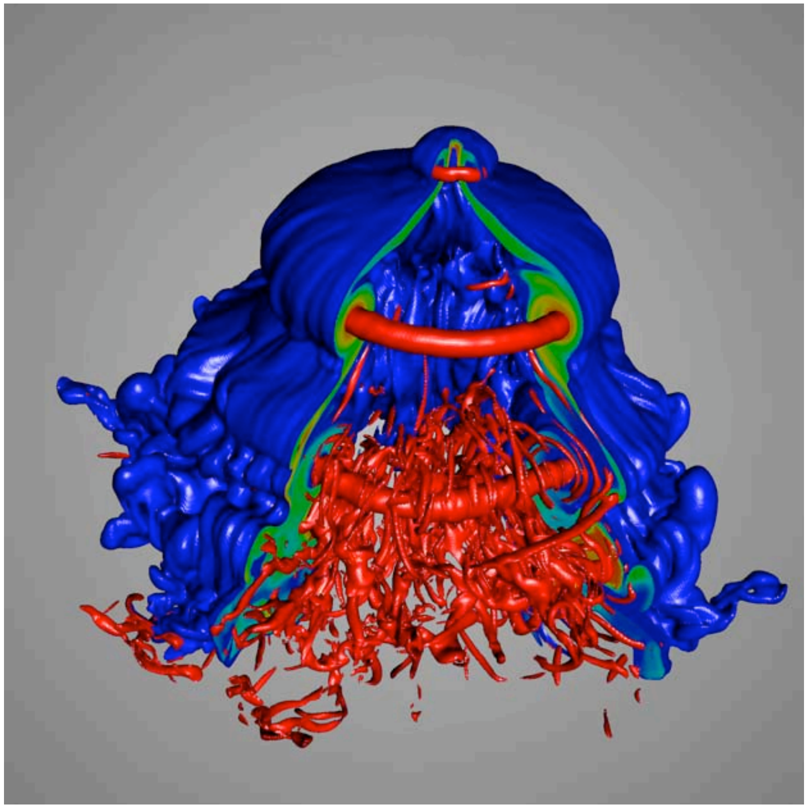


Instability and Turbulence - Raptor

• Multi-physics Eulerian Adaptive Mesh Refinement (AMR) code used for applications at LLNL including astrophysics, Inertial Confinement Fusion (ICF) and shock-driven instabilities and turbulence

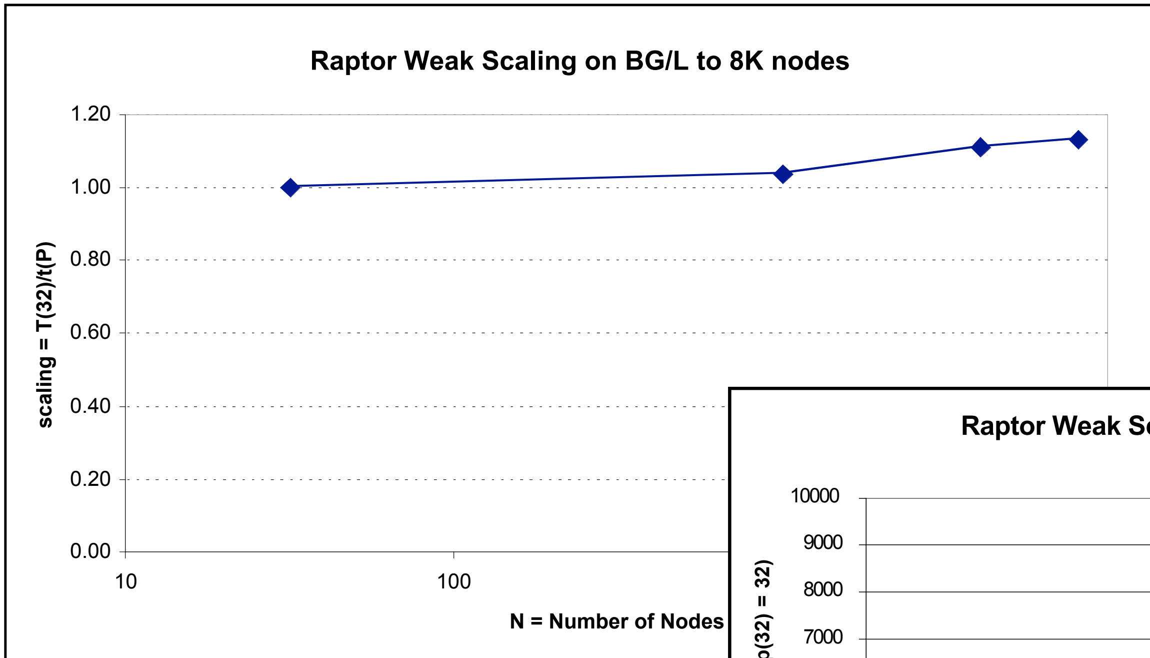
• Can be used to simulate purely fluid dynamics systems and more complex physical systems where the fluids are coupled to the radiation field, such as in ICF or astrophysics

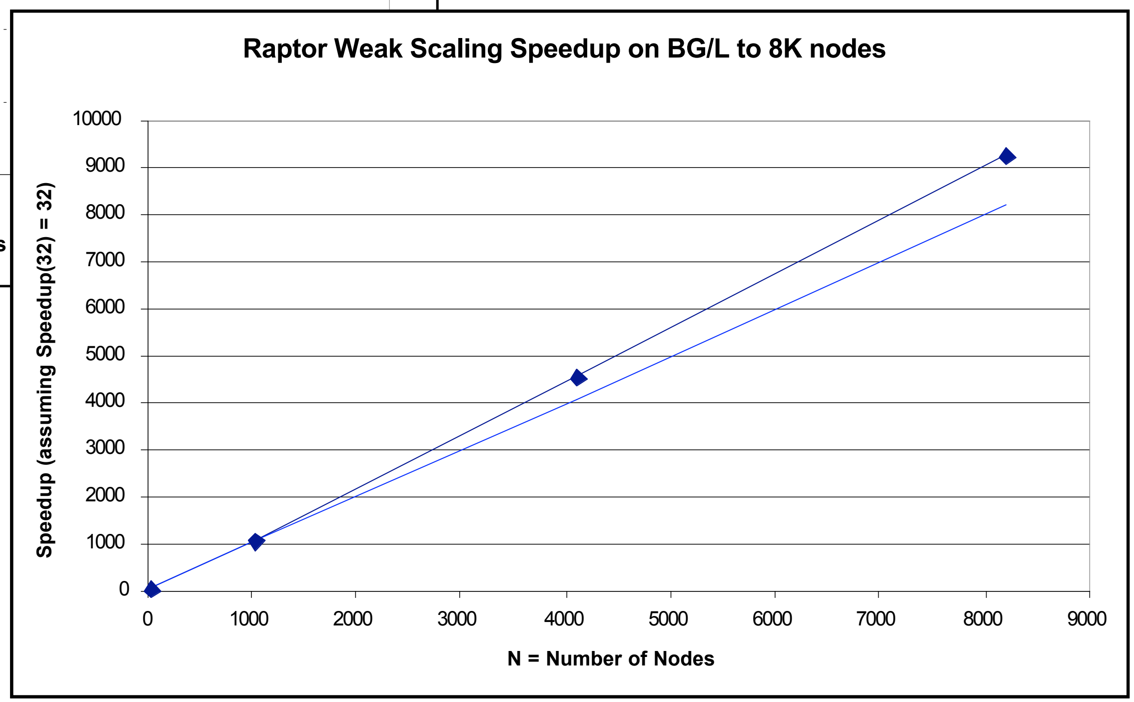
Simulations at full scale on BG/L will offer the computational power to gain an order of magnitude more resolution in simulations of three-dimensional shock-driven systems.



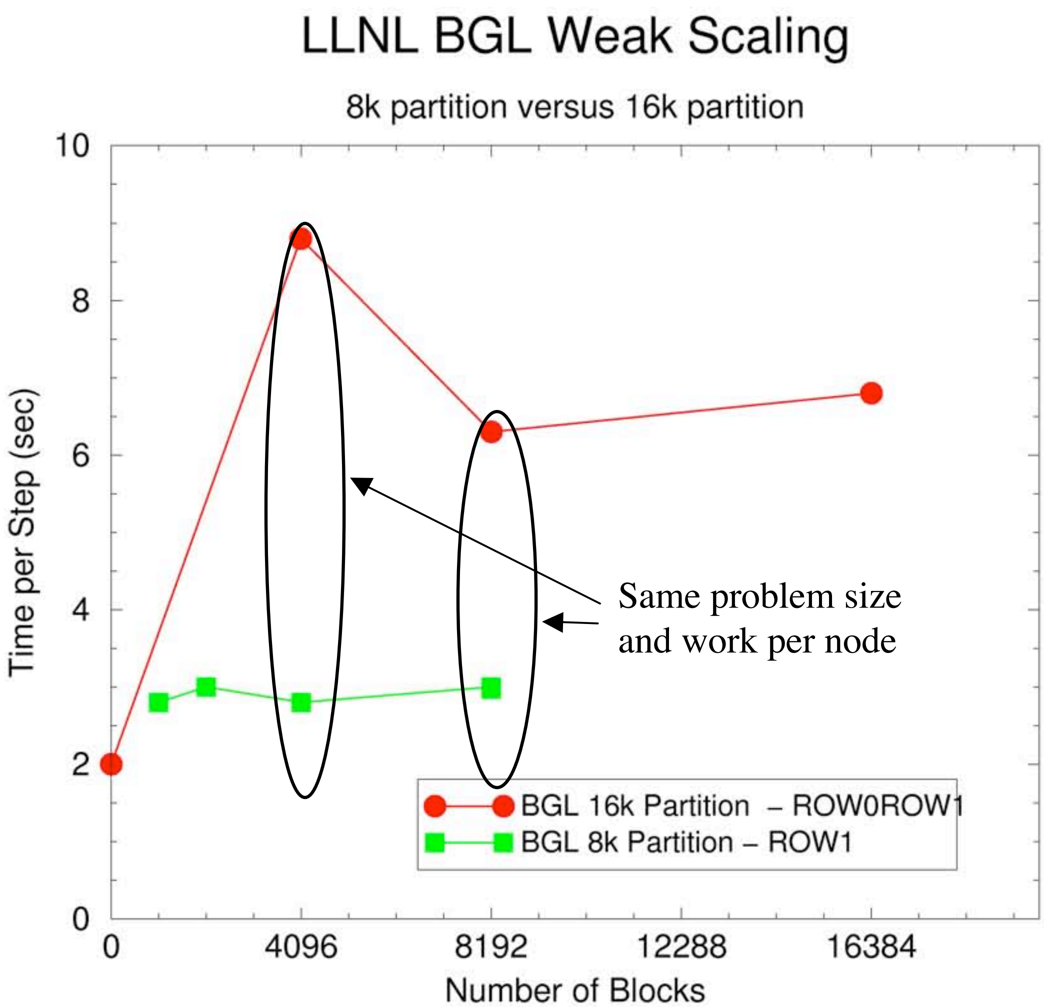
Dense spherical argon bubble, initially contained in a thin spherical soap film, suspended in nitrogen, subjected to a strong planar shock wave about 509 microseconds after shock-bubble interaction. Blue represents the argon iso-surface, red indicates vorticity magnitude, and the film material volume fraction is plotted on the cross-sectional cut planes.

Good scaling results seen for Raptor (note better than linear weak scaling)





Current Raptor Scaling Efforts Focus on 16K Performance Anomaly



- Current efforts aimed at understanding differences between 8k and 16k performance.

- A communications kernel that simulations the point-to-point message “storms” in Raptor has been written and tested.

Planned Multi-Physics Runs - ALE3D (not yet running, export control issue)

Explosives Applications

Using the capability of ALE3D and the resources of BG/L will simulate processes by which chemical explosives detonate (from intended or accidental stimuli) to develop safer explosives

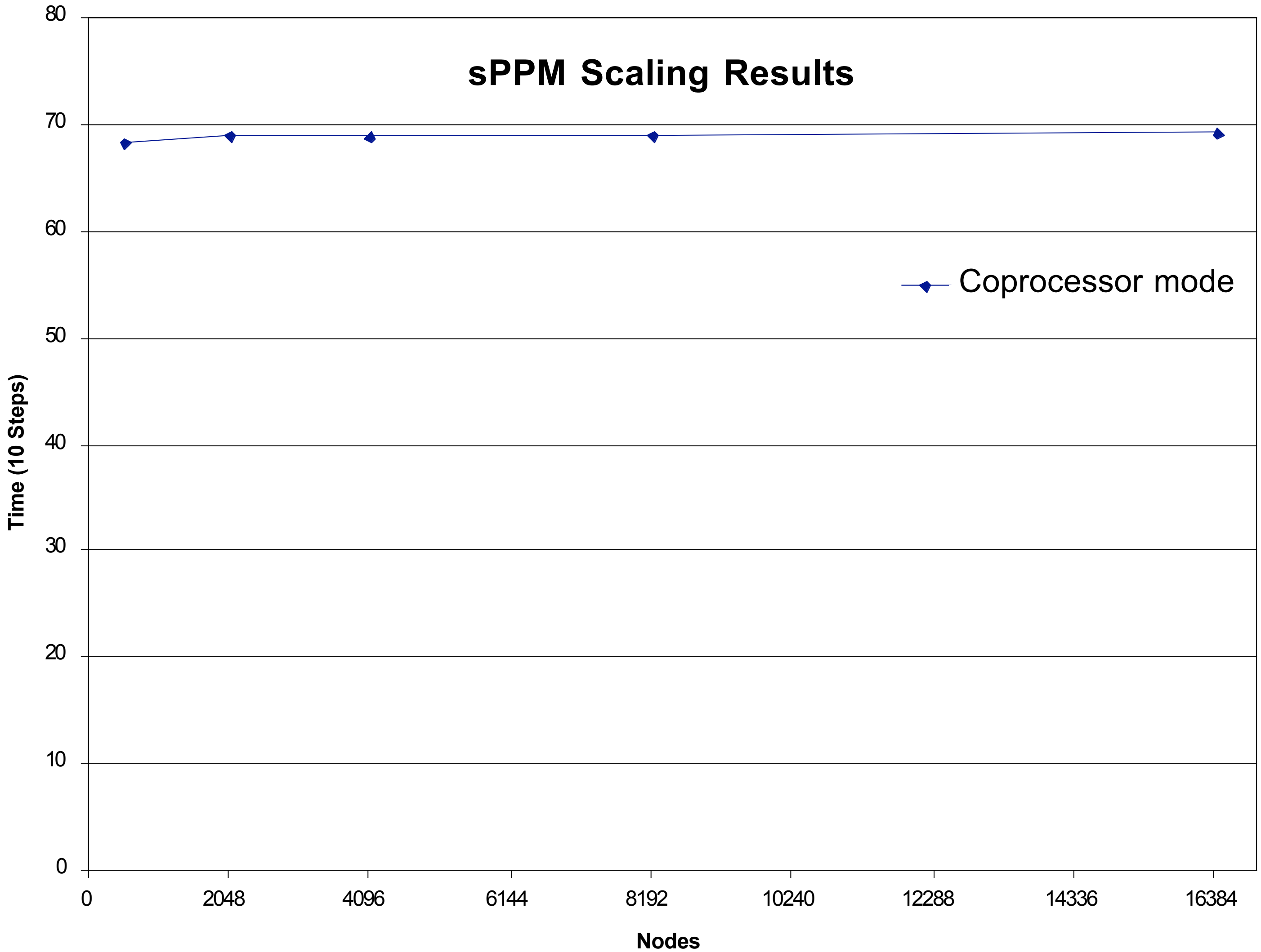
Poly-crystal Plasticity

Fragmenting cylinder calculations in ALE3D will be the first full-scale simulations using material behavior determined directly from a poly-crystal plasticity model

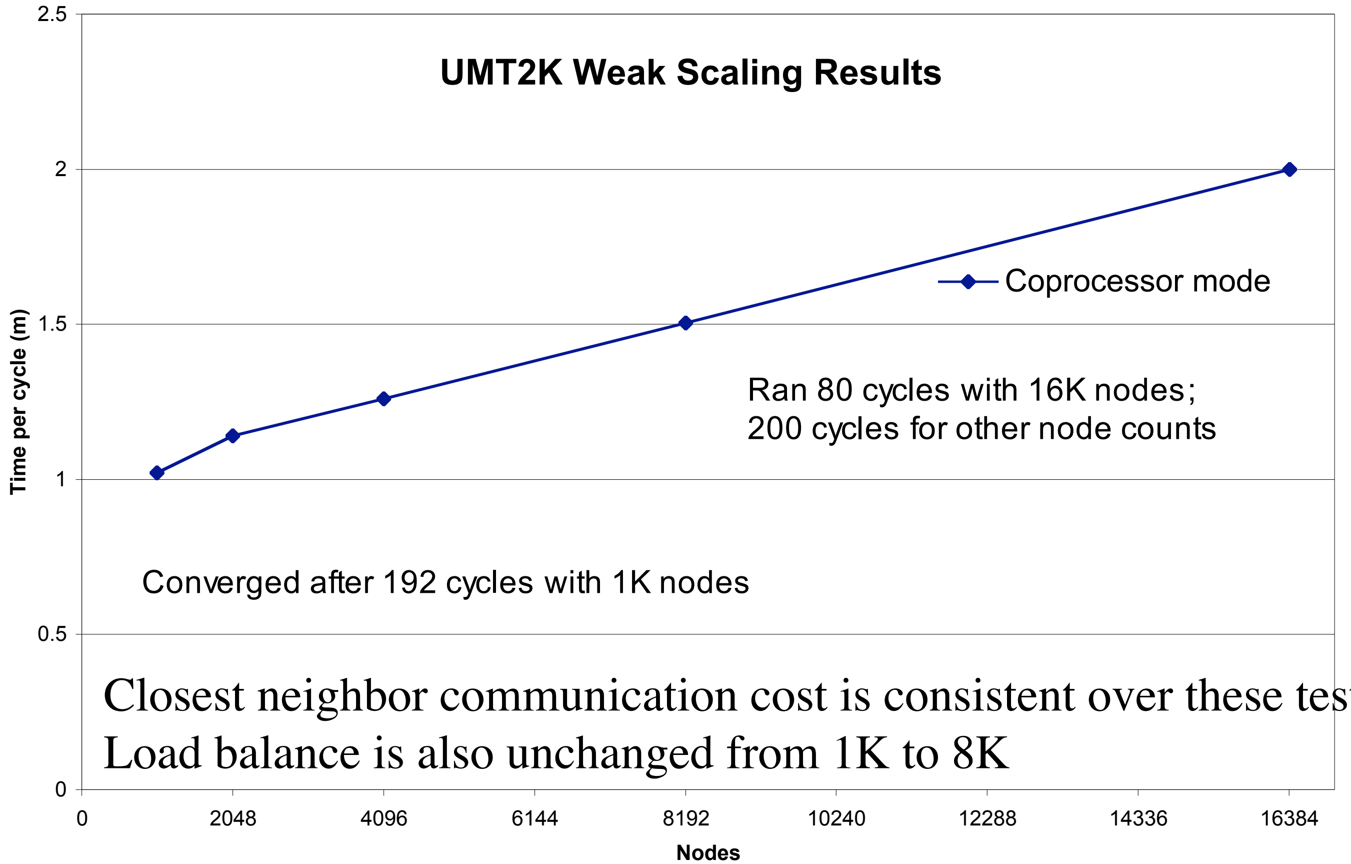
Military Thermal Safety

With ALE3D and BG/L, LLNL will be the first research lab to simulate cookoff in the various phases of high-explosive heating and phase change

sPPM demonstrates near perfect weak scaling to 16,384 nodes



Recent UMT2K runs demonstrate good performance up to 8192 nodes



Closest neighbor communication cost is consistent over these test;
Load balance is also unchanged from 1K to 8K

Virtual node provides a 1.7X speed-up for small scale tests

SMG2000 is semi-coarsening multigrid method found in the hypre library

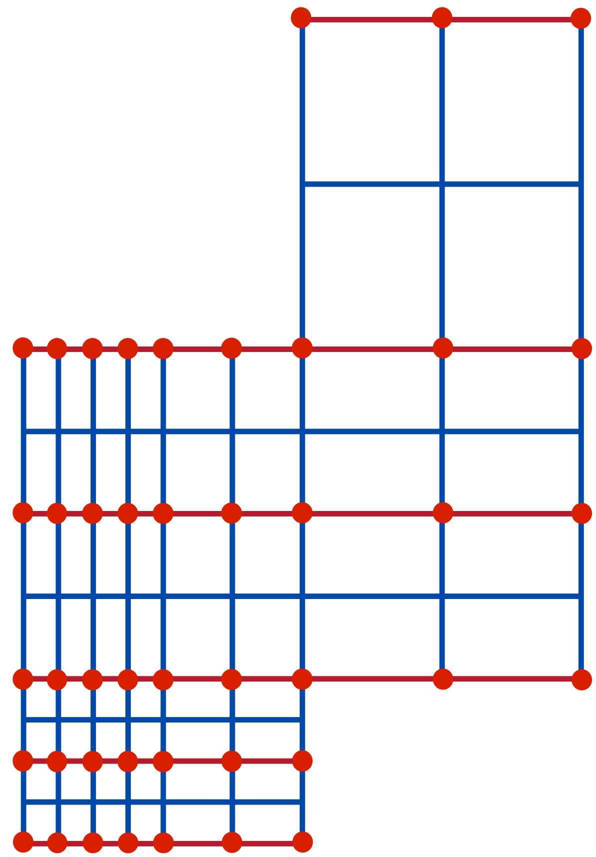
- Robust implementation of a fairly expensive method used in several ASC codes

- 3D SMG calls 2D SMG (to do plane smoothing) calls 1D SMG (to do line smoothing)

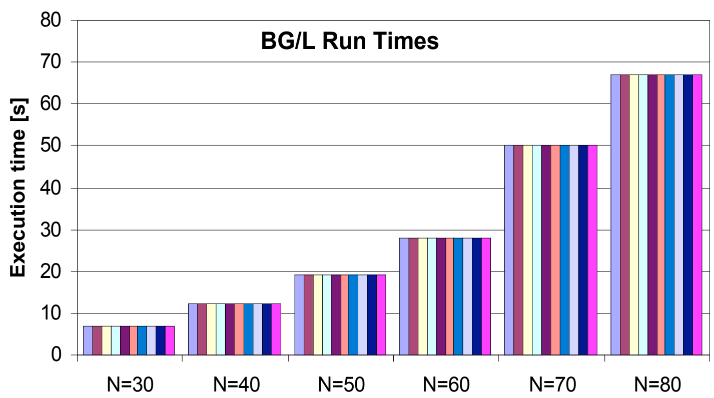
- Parallel model, assuming n^3 data per process and $(pn)^3$ global grid:

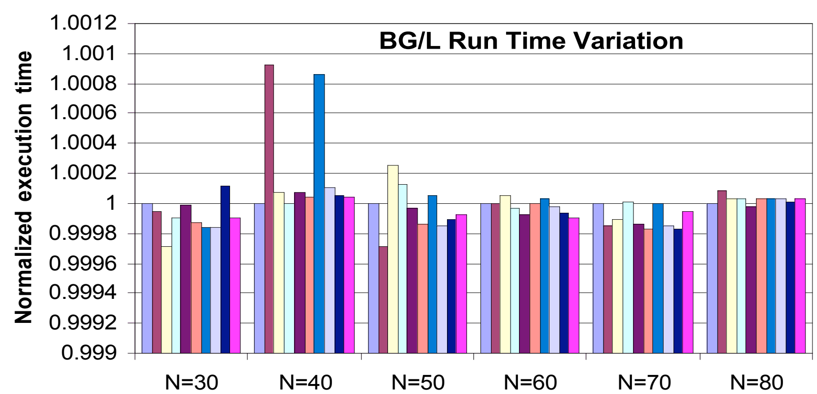
$$T \approx (4L + 8L^2 + 8L^3)\alpha + 20Ln^2\beta + 48n^3\gamma$$

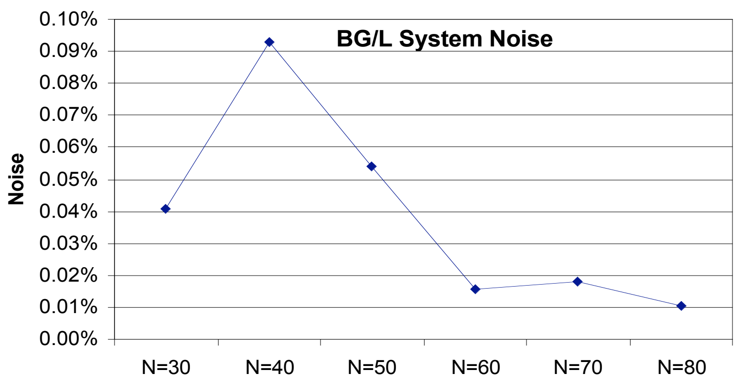
The recursive nature of the algorithm leads to complex communications patterns; good for stressing new architectures; input parameter sensitivity study complete for runs on a single midplane.

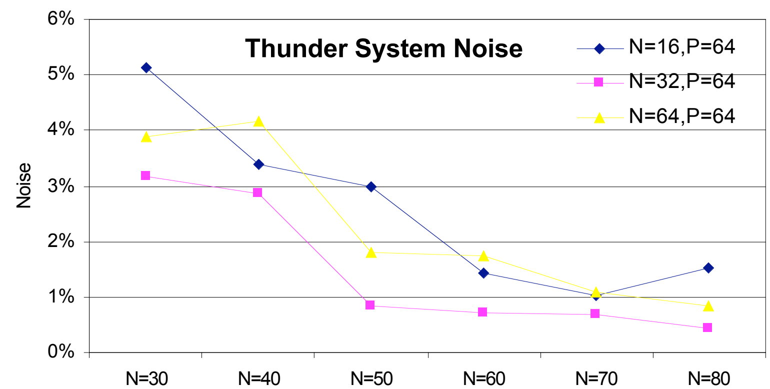


SMG2000 runs demonstrate BG/L's consistent system performance









BlueGene/L promises to revolutionize DOE mission and high-end computing

BG/L is already the fastest computer in the world, at only 1/4 the size of its eventual 64-rack configuration at LLNL this summer…

Linpack numbers and the Top 500 are certainly exciting news events, with IBM, BG/L and DOE at the top once again…

BUT, the application results such as those just presented are what all the excitement should really be about:

- **Enabling better science**
- **Impact on national mission**
- **Cost-effective path to petaFLOP/s**
- **Validating BG/L HW & SW design and capabilities**